Report of Year 2 Activities: Solar Loop Mining to Support Studies of the Coronal Heating Problem (Applied Information Science Research Program: NNG05GB53G)

PI: Olfa Nasraoui, University of Louisville
Co-PI: Joan Schmelz, University of Memphis
Graduate Students: Nurcan Durak, Heba Elgazzar, Sofiane Sellah, Carlos Rojas
Dept of Computer Engineering and Computer Science,
University of Louisville
Visiting Post-docs: Fabio Gonzalez, Jonatan Gomez

ABSTRACT

The Coronal Heating Problem is one of the longest standing unsolved mysteries in astrophysics. Measurements of the temperature distribution along the loop length can be used to support or eliminate various classes of coronal temperature models. The temperature analysis of coronal loops is a state-of-the-art astronomy. In order to make progress, scientific analysis requires data observed by instruments such as EIT, TRACE, and SXT. The combination of EIT, TRACE, and SXT information provides a powerful data set that will yield unprecedented detail on the plasma parameters of a variety of coronal loop structures. The biggest obstacle to completing this project is putting the data set together. The search for interesting images (with coronal loops) is by far the most time consuming aspect of this project. Currently, this process is performed manually, and is therefore extremely tedious, and hinders the progress of science in this field. We propose an approach based on data mining to quickly sift through massive data sets downloaded from the online NASA solar image databases and automatically discover the rare but interesting images with solar loops, which are essential in studies of the Coronal Heating Problem. The proposed solar loop mining scheme will rely on the following components: (i) Collection and labeling of a sample data set of images coming from both categories (with and without solar loops), (ii) An optimal feature selection strategy that will facilitate the retrieval task, (iii) A classification strategy to classify the transformed image into the correct class, and (iv) Appropriate measures to validate the effectiveness of the loop mining process. This project will be implemented in three main phases that target the image databases collected by two different instruments, EIT aboard the NASA/European Space Agency spacecraft SOHO and NASAs TRACE. We will leave open the possibility of targeting the SXT database on the Japanese Yohkoh spacecraft if time permits. All the results of this project: literature, software, and mined Semantic loop features and class labels (in ASCII and XML formats) on tested portions of the different instrument databases will be made available to the public and other interested researchers via the World Wide Web.

1. Highlights

- In Year 1, we have started with generic loops that are easily detected by non-experts, while waiting for the expert labels to be provided.
- In Year 2, we have re-coded all the Matlab code in Java, so that the entire code can be integrated, installed and ported easily on any platform, and can be used as freeware.
- In Year 2, our focus moved to the much harder, rarer, and special category of solar loops in EIT images as determined by the labeled training data that has finally been provided by our collaborators from the University of Memphis.
- We have investigated different pre-processing and data mining options, and tested different combinations of feature sets.
- We constructed additional features, and in addition performed an extensive number of experiments with different classifiers, as well as different cost matrices, and different training sample selection strategies and learning approaches, such as boosting.

- We have isolated data according to its solar cycle, and performed experiments that compare results of different solar cycles versus combined cycles.
- It was discovered that some of the labeling was inconsistent because of subjectivity of different subjects. For example, some loops were labeled as no-loops.
- These labeling problems made it very difficult to design good classifiers no matter which pre-processing, features, or classification method was used.
- Several classification algorithms were compared, including Ripper (a propositional rule learner that learns very concise rules, using Repeated Incremental Pruning to Produce Error Reduction) [16], C4.5 decision trees [20], Multi Layer Perceptron Neural networks (MLP), Support Vector Machines (SVM) [21], Naive Bayes classifiers, and Adaboost [22]. Some of the classifiers are listed in Table 1

Table 1: Some of the investigated classifiers

Classifier	Abbreviation	Brief description
Adaptive Boosting	AdaBoost	Boosting algorithm that learns an ensemble of C4.5 base learners that gradually focus on examples that are hard to classify
Support Vector	SVM	AKernel based method that learns an optimal decision boundary in a higher dimensional projected space.
Naïve Bayes	NB	Probabilistic (Bayesian) classifier
Multilayer Perceptron	MLP	Neural Network Classifier trained using backpropagation
C4.5 Decision trees	C4.5	Decision tree method that learns a tree based classifier built with the most predictive attributes
RIPPER	RIPPER	Learns an optimal set of rules that cover the training samples
K-nearest neighbor	K-NN	Lazy Instance Based classifier

Table 2: Block-based 10-fold cross-validation using low level features

Classifier	**Precision**	**Recall**
AdaBoost_C4.5	0.482	0.282
SVM	0.476	0.115
C4.5	0.523	0.468
RIPPER	0.557	0.518
Naive Bayes	**0.453**	**0.725**
K-NN	0.366	0.349
Multilayer perceptron	0.598	0.482

Table 3: Block-based 10-fold cross-validation using edge based features

Classifier	**Precision**	**Recall**
AdaBoost_C4.5	0.427	0.376
SVM	0.547	0.216
C4.5	0.519	0.44
RIPPER	0.5	0.557
Naive Bayes	**0.522**	**0.677**
K-NN	0.382	0.372
Multilayer perceptron	0.537	0.415

Table 4: Block-based 10-fold cross-validation using combined features

Classifier	Precision	Recall
AdaBoost_C4.5	0.397	0.328
SVM	0.538	0.209
C4.5	0.523	0.44
RIPPER	0.519	0.511
Naive Bayes	**0.498**	**0.706**
K-NN	0.382	0.378
Multilayer perceptron	0.475	0.342

- Hence, we have re-examined the labels for each training instance, and made the necessary corrections. This was done by involving both the Memphis as well as the Louisville team.
- To accelerate the label verification and correction, we have designed an interactive user-friendly interface that can quickly toggle between block and full image views to view the blocks in context.
- Furthermore, because a given solar image can contain different regions with different solar phenomena (loops, eruptions, flares, or no activity, etc), it was necessary to generate the training data by using each image to generate *several* training blocks. Because of the high number of blocks (each image can generate on the order of 20 to 30 blocks), it was out of question to label each block manually; rather an automatic (though less perfect option) was necessary.

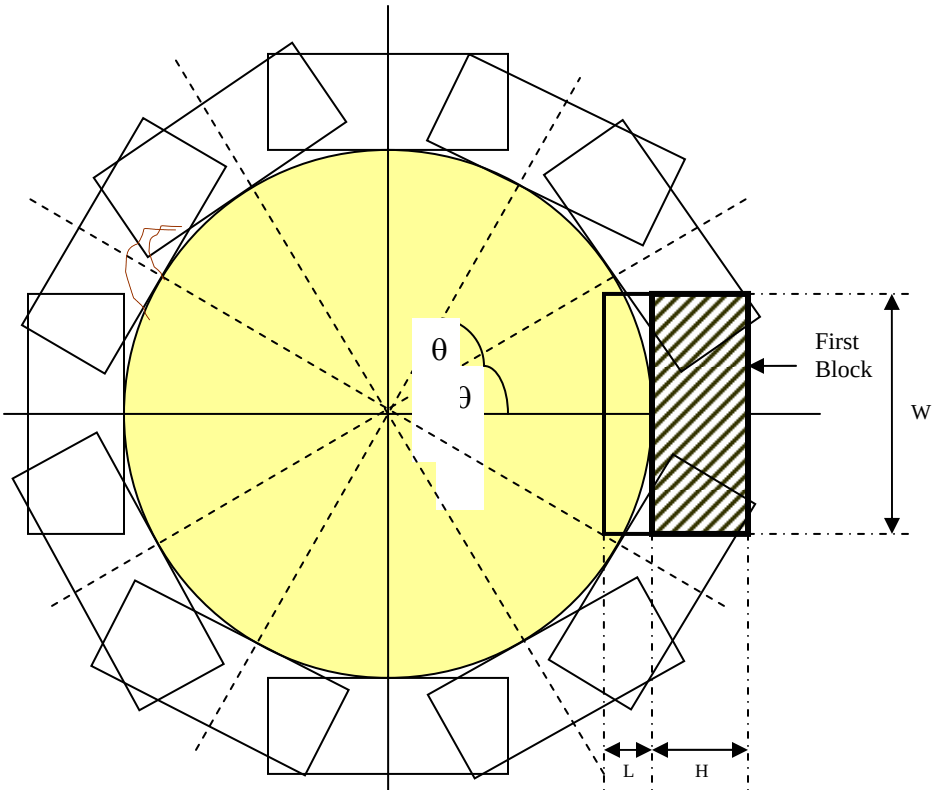


Figure 1: Example showing blocks outside the Solar disk

- These blocks are therefore labeled automatically based on their degree of intersection with an area of interest (that contains a loop) and that has been marked by the labeling subject.
- We have found that different subjects occasionally outlined a loop into regions of varying sizes. It was essential that a region enclosed only the loop and not much other space. Otherwise our automated block extraction and labeling process would result in miss-labeled blocks. This required re-marking several regions of interest so that they satisfy the above requirement.
- After re-labeling and re-marking the training data, we have repeated all the pre-processing and classification experiments.
- We have also constructed and experimented with different features that can capture the loops, and have divided the features into two sets: low level and high level features. The former captures general intensity properties, while the second set captures edge properties and are extracted based on the Hough transform of the edge image as well as EHD (Edge Histogram Descriptors), and Gabor texture features. We have found that the Hough based features that are extracted from the edges gave the best results, and that the other features added little improvement.
- We noticed the presence of not only one kind of loop but several kinds, in particular, some very weak or faint loops were not distinguishable from the background (even after pre-processing the images). These weak loops caused a confusion between the 2 classes while training classifiers. As a result, we divided the loops into 2 classes: loops and weak loops. This resulted in a 3 class problem: loop, weak loop, and no loop, that we felt would make it easier to focus on the subtle differences between loops and non loops.

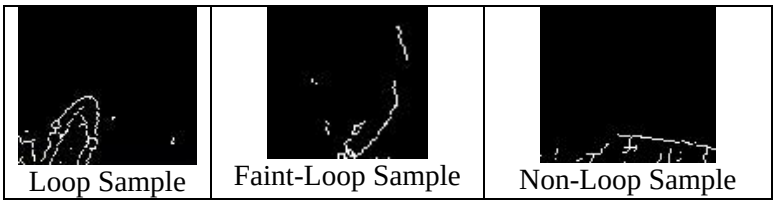


Figure 2: Block samples from different classes

- Results without distinguishing between stronger loops and faint loops are shown in Table 5

Table 5: Block-based 10-fold cross-validation Precision/Recall Results in Loop Class using High Level Features + Hough Line (Hough Line was applied on binary blocks) when training with 2 classes

Precision	Recall	Classifier
0.571	0.59	**AdaBoost_C4.5**
0.607	0.481	**SVM**
0.621	**0.67**	**C4.5**
0.599	**0.691**	**RIPPER**
0.562	0.63	**Naive Bayes**

- Faint loop blocks can be more quickly identified by scrutinizing the missed loops from a first stage of classification using the original 2 class labels (loop and no-loop)
- Results after taking into account faint loops:

Table 6: Block-based 10-fold cross-validation Precision/Recall Results in Loop Class using High Level Features + Hough Line (when faint loops are isolated in a 3rd class)

Precision	Recall	Classifier
0.638	0.62	**AdaBoost_C4.5**
0.669	0.533	**SVM**
0.659	0.666	**C4.5**
0.635	**0.738**	**RIPPER**
0.6	0.668	**Naive Bayes**

- Adding another class can make it easier for the classifiers to "learn" what truly distinguishes one class from another
- In addition to exploring different "kinds" of features, we have investigated ways to attach *location* information to the features, for instance by calculating the features separately in several different bands that differ by their altitude above the Solar photosphere. Our best result was for splitting each block into 4 bands, and the best results were obtained when edge-based features were extracted from the top band (the furthest from the photosphere). This is likely due to the noise the abundance of false (i.e. not coronal loops) phenomena located near the photosphere. Intensity-based features did not seem to be location sensitive.
- To summarize our best performing features so far include (1) the following low level features that were extracted from the intensity levels of the pixels in each block: Mean, Standard Deviation, Smoothness, Third Moment, Uniformity, and Entropy; (2) high level features that were extracted from the binary edges of the image: Line Direction features (based on angles, as explained below), number of edge pixels, number of line segments, length of the longest line segment (crude estimate of number of edge pixels on estimated lines), and number of edge pixels in each of 4 horizontal bands that make up each block. The latter takes into account location information. Lines were estimated using a crude Hough transform. From our preliminary experiments, the well known Edge Histogram Descriptor features or EHD have resulted in worse results than our Line Direction features (based on slope angles of the crudely estimated lines). For Line Direction features, we have distinguished between different directions of linear segments based on their estimated slopes (from the Hough space), by mapping the ranges of slopes in degrees, as shown below, with angles located outside these intervals considered as non-directional.
    * Horizontal: [0,20] and [160,180].
    * Vertical: [70,110].
    * Diagonal: [35,55] and [125,145].

Table 7: 2-class Block-based 10-fold cross-validation Precision/Recall Results in Loop Class using High Level Features + Hough Line with location sensitive features computed on 4 bands (Band 1 = furthest from photosphere, …, Band 4 = adjacent to photoshere): All of the High Level Features except Third and Fourth Band: HorizontalEdges, VerticalEdges, DiagonalEdges, Non-Directional Edge Histogram, Number of Edge Pixels, Number of points on longest straight line (maximum HT accumulator value), Number of Edge Pixels in Band 1, Number of Edge Pixels in Band 2. Underlined features are location-dependent.

Precision	Recall	Classifier
0.649	0.623	**AdaBoost_C4.5**
0.714	0.36	**SVM**
0.662	0.712	**C4.5**
0.631	**0.784**	**RIPPER**
0.628	0.698	**Naive Bayes**

- In general, loop blocks can be very diverse in their shape, size, and direction, and in some cases, are very hard to distinguish (even to the untrained Human eye) from other solar phenomena that occur in the corona, such a coronal mass ejections and solar flares. Some difficult blocks are classified as no-loop by most classifiers because they share a lot of similarity with no-loop blocks. Fuzzy sets can help model these loop blocks better. Figure 4 shows a few examples of loop blocks that are hard to classify.

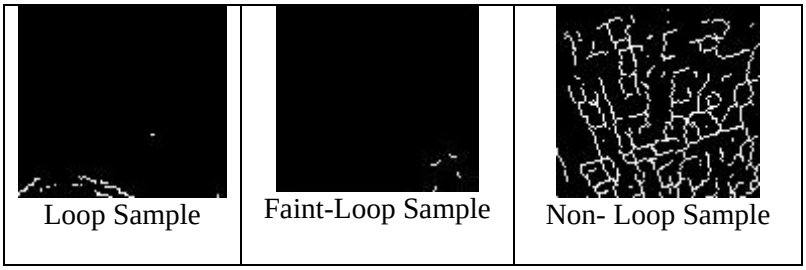


Figure 3: Sample blocks which are hard to classify, showing an example of a block with a faint loop

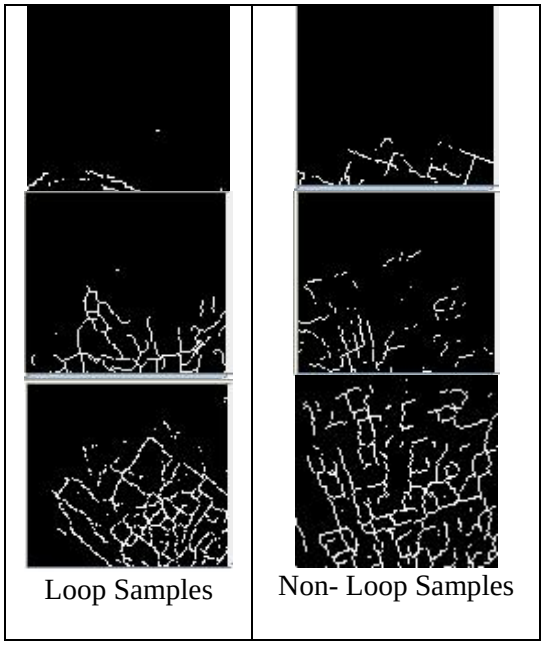


Figure 4: Sample blocks which are hard to classify correctly

- Modeling both input features and output labels using fuzzy set membership values can help provide a more accurate representation of the ground truth and may help in classification of borderline blocks. In order to construct the fuzzy rules, we have first selected the optimal feature subset from RIPPER's rules and the nodes in the C4.5 decision tree. These features were: the number of edge pixels, the number of lines segments, the number of edge pixels in the top band, and the length of the longest line. The output of the fuzzy inference system is the block label which can be loop or no-loop. We determined the value ranges of the input parameters from the ranges in RIPPER's rules and the rules derived from the branches of the C4.5 decision tree. Thus, for each

input parameter, we have defined three intervals: low, medium, and high. Some features had an additional value (TooHigh).

- Mamdani's fuzzy inference systems consist of if-then rules in the form "*If x is A then y is B*", where x and y are fuzzy variables, and *A* and *B* are fuzzy values. We selected the fuzzy rules from the rules that were generated by RIPPER and from short and pure branches in the tree generated by C4.5 decision tree learning. The resulting fuzzy rules are assigned weights that are used to combine all the outputs into the final decision. The weights were defined based on the number of data correctly classified in the corresponding RIPPER and C4.5 rules.

Features used	**Low Level Features**		**High Level Features**	
Classifier	**Precision**	**Recall**	**Precision**	**Recall**
AdaBoost_C4.5	0.537	0.393	0.638	0.628
SVM	1	0.002	0.719	0.37
C4.5	0.538	0.416	0.653	0.695
RIPPER	0.54	0.36	**0.641**	**0.763**
Naive Bayes	0.305	0.158	0.641	0.709
MultiLayerPerception	0.57	0.37	0.644	0.693
Mamdani Fuzzy Rules constructed with RIPPER and C4.5	-	-	**0.659**	**0.781**

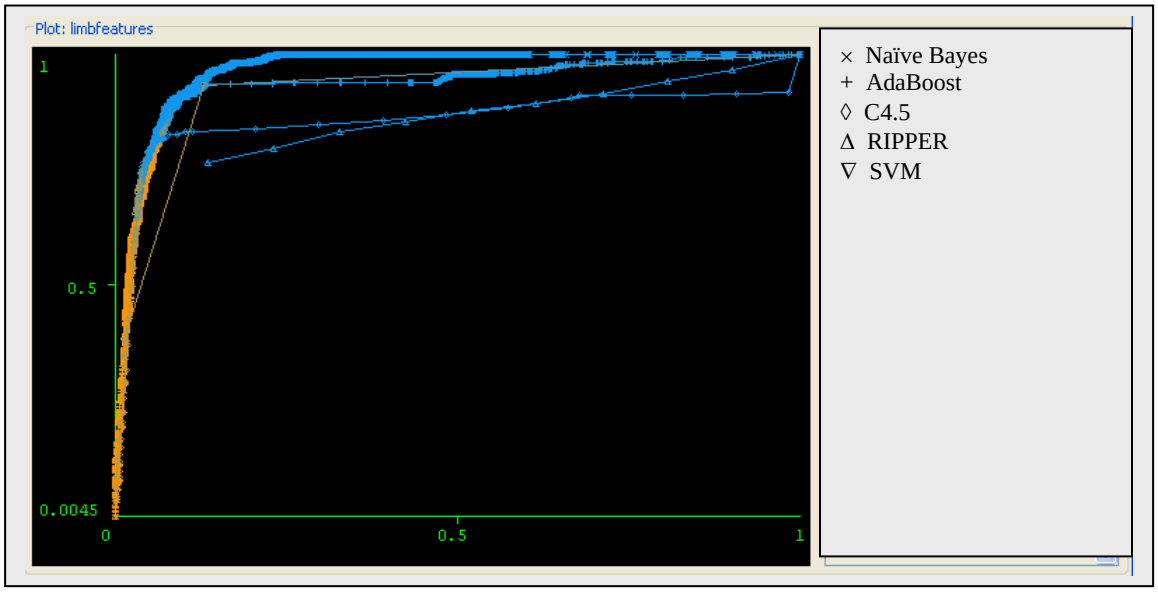
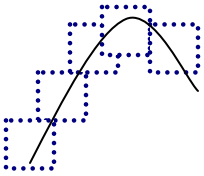


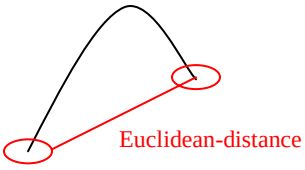
Figure 5: ROC curve for best classifiers for best performing features for out-of-disk loop detection

- After tuning our results as explained above, we have once again turned to investigate more features that can capture the concept of a loop. This has led us to the following new features which are based on "curvature" of the edges:

  1. *Average tangent difference:* After tracing edges in a block, the curvature of the *longest edge* was calculated by sliding a window along the edge, and computing the average on all windows (i) of the differences between consecutive window edge tangent angles (Θ_{i+1} - Θ_i) or slopes ($\Theta_i=(y_2-y_1)/(x_2-x_1)$) at the extreme points (x_1,y_1) and $(x_2,y_{2)}$ in each window.



2. *Length of Curvature Edge:* This is the length of the longest edge in a block which is calculated after edge tracing.
3. *Sign Change distribution:* based on the number of window angles that are below 0^o, N_-, and Number of window angles that are above 0^o, N_+, as follows:
   Sign-change = min (N_+, N_-) / max (N_+, N_-)
4. *Euclidean Distance between two endpoints of the longest edge:*



5. *Alternative_Curvature* = (Distance between two endpoints) / (Curvature length)

Several experiments were performed with different attribute combinations. The best results are shown below. In these results, RIPPER gave slightly better result in recall than the best results so far, while maintaining a similar precision level. Figure 5 and 6 show the ROC curve and the Precision versus Recall curve, respectively for the best performing features and algorithms so far.

Table 8: 2-class Block-based 10-fold cross-validation Precision/Recall Results in Loop Class using the *best* subset of high level features which contains the following attributes: L_1=Length of Longest Curved Line, L_2= Euclidean Distance Between End Points of Longest Curved Line, L_3= L_1 / L_2, Non-Directional Edge Histogram, Number of Edge Pixels, Number of points on longest straight line (maximum HT accumulator value), Number of Edge Pixels in Band 1, Number of Edge Pixels in Band 2. Underlined features are curvature-related.

Precision	Recall	Classifier
0.637	0.616	**AdaBoost_C4.5**
0.708	0.36	**SVM**
0.647	0.751	**C4.5**
0.631	**0.804**	**RIPPER**
0.593	0.721	**Naive Bayes**

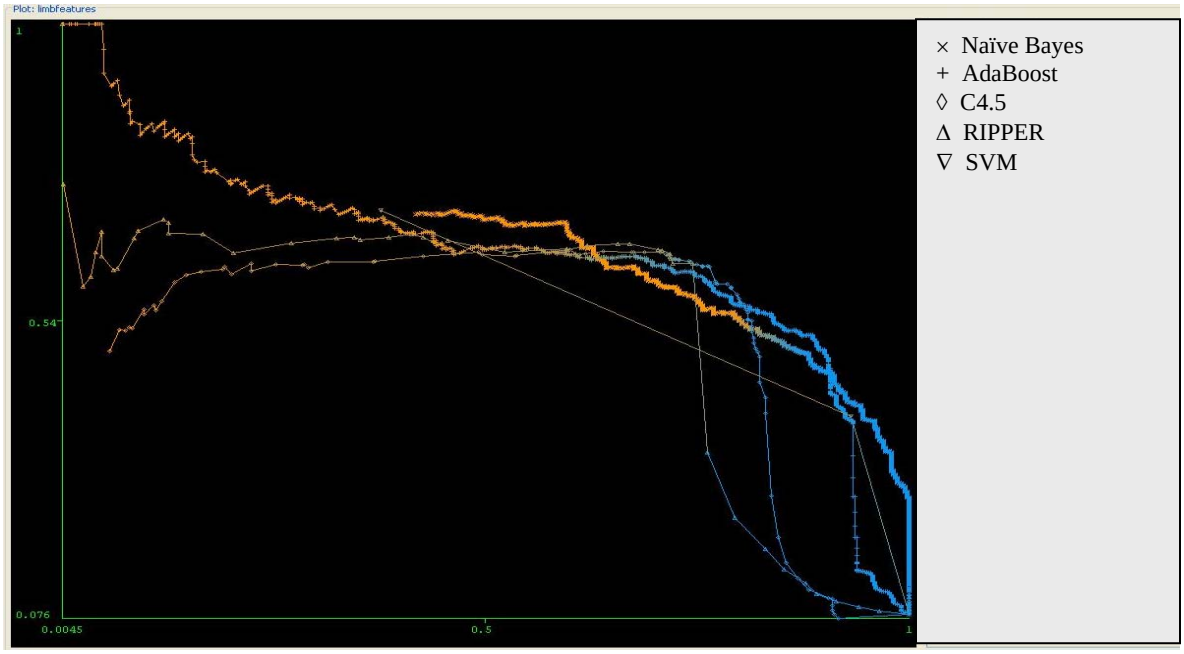


Figure 6: Precision versus Recall curve for best classifiers for best performing features for out-of-disk loop detection

- We have also started classification of loops located *within* the Solar disk, and will report our results in the near future.

2. Relevance to NASA

The search for interesting images for coronal temperature analysis (with coronal loops) amounts to searching for a needle in a haystack, and therefore hinders the fast progress of science in this field. The next generation EIT called MAGRITE, scheduled for launch in a few years on NASA's Solar Dynamics Observatory, will require state of the art techniques to sift through the massive wealth of data to support scientific discoveries. The proposed work addresses goals 1 and 2 of the Applied Information Systems Research (AISR) program, since it includes novel information technology and computational methods that promise to increase productivity of the OSS research and public outreach endeavors, and would benefit the state-of-practice in space science. It also fosters interdisciplinary collaboration spanning the space science (Co-I) and computer science (PI) disciplines. Our project addresses objective 4 of the AISR Program, namely increasing science and educational return from the data through advanced knowledge discovery methodologies.

3. Application to NASA Missions and Programs

The Coronal Heating Problem is one of the longest standing unsolved mysteries in astrophysics. Measurements of the temperature distribution along the loop length can be used to support or eliminate various classes of coronal temperature models. The temperature analysis of coronal loops is a state-of-the-art astronomy. In order to make progress, scientific analysis requires data observed by instruments such as EIT, TRACE, and SXT. The combination of EIT, TRACE, and SXT information provides a powerful data set that will yield unprecedented detail on the plasma parameters of a variety of coronal loop structures. The biggest obstacle to completing this project

is putting the data set together. The search for interesting images (with coronal loops) is by far the most time consuming aspect of this project. Currently, this process is performed manually, and is therefore extremely tedious, and hinders the progress of science in this field. Our project aims to accelerate and automate the discovery of the rare but interesting images with solar loops.

In addition to the specific problem from Astrophysics, above, research that advances state of the art in solar physics will have a significant impact on society and other scientific fields because of the following reasons: (i) The climate connection: the sun is a source of light and heat for life on Earth. Scientists strive to understand how it works, why it changes, and how these changes influence the Earth, (ii) Space weather: The sun is the source of the solar wind: flow of gases from the sun that streams past the Earth at speeds exceeding a million miles per hour. Disturbances in the solar wind shake the Earth's magnetic field. Space weather can change the orbits of satellites and shorten mission lifetimes. Excess radiation can physically damage satellites and poses a threat to astronauts, in addition to power surges and outages on Earth, and hence needs to be predicted. (iii) The sun as a physical laboratory: the sun produces its energy by nuclear fusion, a process that scientists have strived for decades to reproduce by involving hot plasmas in strong magnetic fields. Much of solar astronomy involves observing and understanding plasmas under similar conditions.

4. Tracking

All the results of this project: literature, software, and outputs (labels) of the developed classification methods on tested portions of the different instrument databases will be made available to the public and other interested researchers via the World Wide Web. Outputs of our automated retrieval process (on new test data) will be saved in both ASCII column format and XML format to facilitate data interchange with and between different research groups. The XML schema will include derived Semantic features (Estimated number of loops and their confidence) and the assigned class labels.

Tracking the usage of our products is easily accomplished by monitoring the access statistics on the projects website, a feature that is already in use, as well as searching for citations on the Web.

5. Software and Publications

Software on the project collaboration platform website:
"http://webmining.spd.louisville.edu/twiki/bin/view/SOLARLoops/".

Publications & Presentations:

- N. Durak, O. Nasraoui, J. Gomez, F. Gonzalez, H. Elgazzar, S. Sellah, C. Rojas, J. Schmelz, J. Roames, K. Nasraoui., “Mining Coronal Loops in Solar Images from the SOHO collection”, In Proceedings of NAFIPS 2007, San Diego, CA, 2007.
- O. Nasraoui and C. Rojas, “Robust Clustering for Tracking Noisy Evolving Data Streams,” in Proc. SIAM conference on Data Mining, Bethesda, MD, Apr. 2006, 618-622.
- O. Nasraoui, H. Elgazzar, C. Rojas, Fabio Gonzalez, Jonatan Gomez , J. Schmelz, and J. Roames, Kaouther Nasraoui, Lindsey Lippner, Jennifer Garst, Andrew Gibson, “Using Data Mining for Automatic Retrieval of Solar Loop Images”. Conf. on Statistical Challenges in Modern Astronomy, Pennsylvania State University, June. 2006.

6. Upcoming Plans

- Publishing the software and mining results of Phase 1: Mining loops located outside the Solar disk in EIT data
- Completing Phase 2: Mining loops inside the Solar disk in EIT data
- Starting Phase 3: Mining loops in TRACE data.
- Collection and labeling of relevant TRACE images for Phase 3 (our collaborators in the Solar Physics lab at the University of Memphis), notably in several categories (low, medium, high level of solar activity) depending on the solar cycle (minimum to maximum)
- EIT data's edge quality is too poor for reliable higher order (e.g. elliptical) fitting with the Hough Transform (HT). However, TRACE data has higher quality edges because it contains close-up high-resolution shots of Solar corona, and should therefore be able to benefit from HT and from *direct* ellipse detection. Thus, several directions are in order for TRACE:
    - Scaling the Hough transformation procedure by incorporating constraints to prune the space of possible parameters in the early stages of computation
    - Integration of the Hough space analysis with single-pass robust clustering (ACRES-Streams) to automate the semantic loop feature extraction
    - Investigating alternative loop detection methods, such as by seeking clusters of elliptically shaped arcs.
- Improving results for Phase 1 and Phase 2 in 2 ways:
    - Building an aggregation decision making system to combine the class outputs of neighboring blocks into a final "image-based" outcome
    - Using a preliminary clustering stage (for all data): Applying clustering (unsupervised categorization) on the varying cycle data (different levels of solar activity) to create homogenous groups of images
    - Developing context-sensitive loop retrieval mechanisms that automatically adapt to the category of each input image